**EXP 4** 210701249

**Create User Define Function in Apache Pig and execute it on map reduce**

**Aim:**

To createUserDefineFunctioninApache Pigandexecuteitonmapreduce

**Procedure:**

1. **Firstly install PIG**

**Step1:**LoginintoUbuntu

**Step 2**: Go to <https://pig.apache.org/releases.html>and copy the path of the latest version ofpigthatyou wantto install.Runthe followingcomment todownloadApachePigin Ubuntu:

$wget<https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz>

**Step3**: Tountar pig-0.16.0.tar.gzfilerunthefollowingcommand:

$tarxvzf pig-0.16.0.tar.gz

**Step 4:** To create a pig folder and move pig-0.16.0 to the pig folder, execute the followingcommand:

$sudomv/home/hdoop/pig-0.16.0 /home/hdoop/pig

**Step 5:** Now open the .bashrc file to edit the path and variables/settings for pig. Run thefollowingcommand:

$sudo nano.bashrc

Addthebelowgiven to.bashrcfile at theend andsavethefile.

#PIG settingsexport PIG\_HOME=/home/hdoop/pigexportPATH=$PATH:$PIG\_HOME/binexportPIG\_CLASSPATH=$PIG\_HOME/conf:$HADOOP\_INSTALL/etc/hadoop/exportPIG\_CONF\_DIR=$PIG\_HOME/confexport JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64exportPIG\_CLASSPATH=$PIG\_CONF\_DIR:$PATH#PIGsettingends

**Step6:**Runthe followingcommandtomakethe changeseffectivein the.bashrcfile:

$source.bashrc

**Step 7:** To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run thefollowingcommands:

$./start-dfs.sh$./start-yarn$jps

**Step8:**Nowyoucan launchpigby executingthefollowingcommand:

$ pig

**Step 9:** Now you are in pig and can perform your desired tasks on pig. You can come out ofthe pig by the quit command:

>quit;

1. **Create UDF in Pig**

**Create a sample text file**hadoop@Ubuntu:~/Documents$ nano sample.txtPastethebelow content to sample.txt

1,John2,Jane3,Joe4,Emma

hadoop@Ubuntu:~/Documents$hadoopfs-putsample.txt/home/hadoop/piginput/

# CreatePIGFile

hadoop@Ubuntu:~/Documents$nanodemo\_pig.pig

# pastethe belowthe contentto demo\_pig.pig

--LoadthedatafromHDFS

data=LOAD'/home/hadoop/piginput/sample.txt'USING PigStorage(',')AS(id:int>

-- Dump the data to check if it was loaded correctlyDUMPdata;

# Run the abovefile

hadoop@Ubuntu:~/Documents$pigdemo\_pig.pig

2024-08-0712:13:08,791[main]INFO

org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil

- Total input paths to process : 1(1,John)

(2,Jane)

(3,Joe)

(4,Emma)

**Createudffileansaveasuppercase\_udf.py**

uppercase\_udf.py

def uppercase(text):returntext.upper()

ifname == "main":import sys

forlinein sys.stdin:

line=line.strip()

result = uppercase(line)print(result)

# Createtheudfsfolderonhadoop

**hadoop@Ubuntu:~/Documents$hadoopfs -mkdir/home/hadoop/udfs**

# puttheupppercase\_udf.pyintotheabvfolder

**hadoop@Ubuntu:~/Documents$hdfsdfs -putuppercase\_udf.py/home/hadoop/udfs/**

# hadoop@Ubuntu:~/Documents$ nano udf\_example.pigcopyand pastethebelowcontentonudf\_example.pig

--Register thePython UDFscript

REGISTER'hdfs:///home/hadoop/udfs/uppercase\_udf.py'USINGjythonASudf;

--Load somedata

data=LOAD'hdfs:///home/hadoop/sample.txt'AS(text:chararray);

--Usethe Python UDF

uppercased\_data=FOREACHdataGENERATEudf.uppercase(text)AS uppercase\_text;

--Storetheresult

STOREuppercased\_dataINTO'hdfs:///home/hadoop/pig\_output\_data';

# placesample.txt file on hadoop

hadoop@Ubuntu:~/Documents$hadoopfs -putsample.txt/home/hadoop/

# ToRunthe pig file

hadoop@Ubuntu:~/Documents$pig -fudf\_example.pig

# finally u getSuccess!

**JobStats(timein seconds):**

JobId Maps Reduces MaxMapTimeMinMapTime AvgMapTime MedianMapTimeMaxReduceTimeMinReduceTime AvgReduceTimeMedianReducetime

AliasFeatureOutputs

job\_local1786848041\_0001 1 0 n/a n/a n/a n/a 00 0 0data,uppercased\_dataMAP\_ONLYhdfs:///home/hadoop/pig\_output\_data,

Input(s):

Successfullyread4records(42778068bytes)from:"hdfs:///home/hadoop/sample.txt"

Output(s):

Successfullystored4records(42777870bytes)in:"hdfs:///home/hadoop/pig\_output\_data"

Counters:

Totalrecords written:4

Totalbyteswritten :42777870

Spillable Memory Manager spill count : 0Totalbags proactively spilled: 0

Totalrecordsproactively spilled:0

JobDAG:

job\_local1786848041\_0001

2024-08-0713:33:04,631[main]WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl -JobTrackermetricssystem alreadyinitialized!

2024-08-0713:33:04,639[main]WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl -JobTrackermetricssystem alreadyinitialized!

2024-08-0713:33:04,644[main]WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl -JobTrackermetricssystem alreadyinitialized!

2024-08-0713:33:04,667[main]INFO

org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -Success!

# Note:

**If any error check jython package is installed and check the path specified on the abovestepsaregivecorrectly**

# --------------------------------------------------------------------------------------------------------------

**Tochecktheoutputfileiscreated**

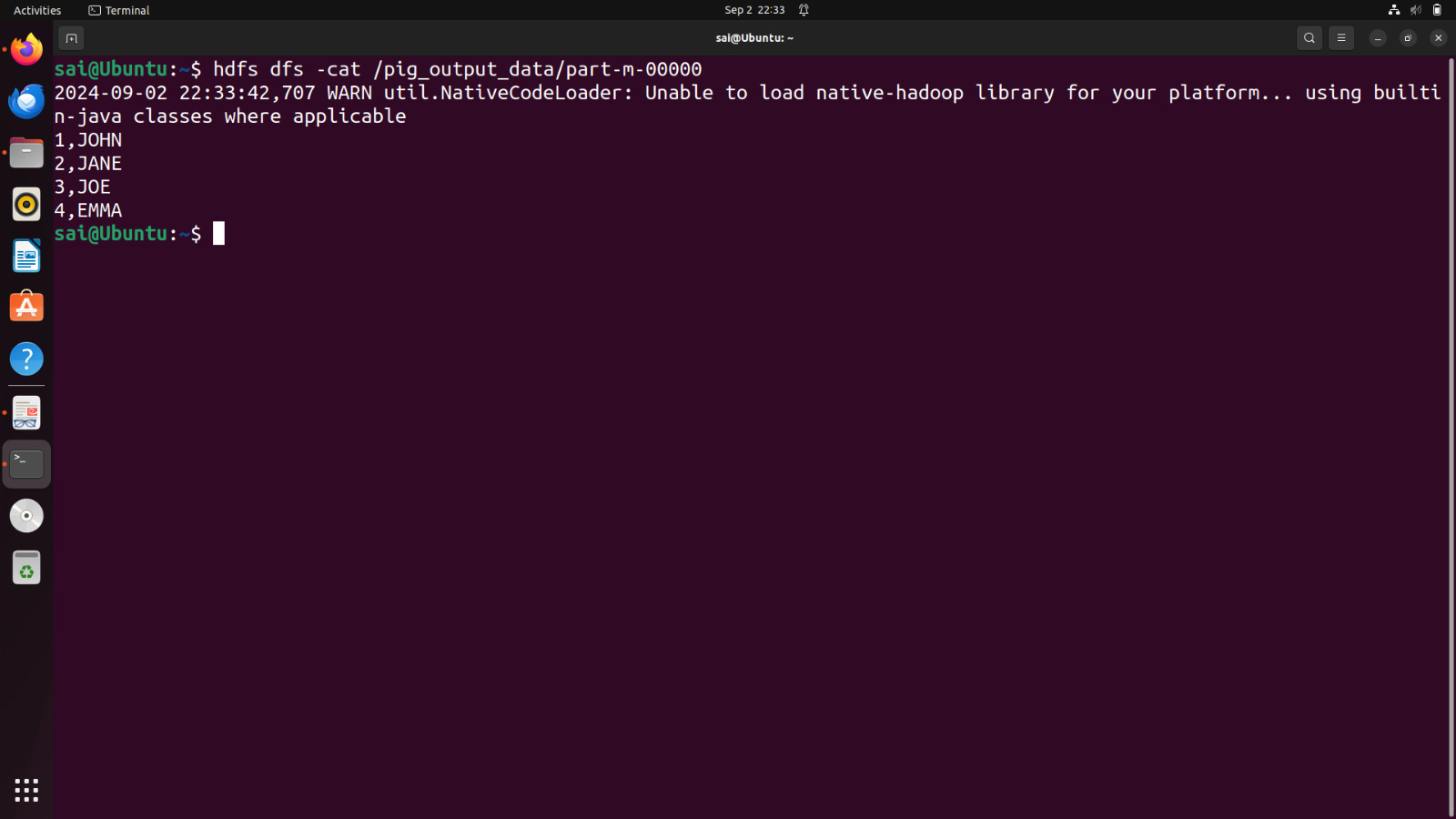
hadoop@Ubuntu:~/Documents$ hdfs dfs -ls /home/hadoop/pig\_output\_dataFound2 items

Ifyouneed toexaminethefilesin theoutput folder,use:

# Toviewthe output

**hadoop@Ubuntu:~/Documents$ hdfs dfs -cat /home/hadoop/pig\_output\_data/part-m-00000**

**OUTPUT:**



**Result:**

Thus theUserDefineFunctioninApache Pigandexecuteitonmapreduce is executed successfully.